Lerneinheit: Manuelle Annotation mit CATMA

Jan Horstmann [1]

1. Universität Münster

forTEXT

Thema:	Manuelle Annotation	DOI:	10.48694/fortext.3750
Jahrgang:	1	Ausgabe:	4
Erscheinungsdatum:	2024-08-07	Erstveröffentlichung:	2019-11-04 auf fortext.net
Lizenz:	CC BY SA		open access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

Eckdaten der Lerneinheit

- Anwendungsbezug: Franz Kafkas *Erstes Leid* (1924)
- Methode: Digitale taxonomiebasierte manuelle Annotation
- Angewendetes Tool: CATMA
- Lernziele: Erstellung eines CATMA-Accounts und eines CATMA-Projektes, Organisation von Ressourcen und Annotation eines Textes mit unterschiedlichen Kategorien
- Dauer der Lerneinheit: ca. 60–90 Minuten
- Schwierigkeitsgrad des Tools: leicht

Bausteine

- Anwendungsbeispiel: Welchen Text annotieren Sie mit welchen Kategorien? Annotieren Sie Kafkas Erzählung *Erstes Leid* anhand der „Distanz"-Kategorien aus (Martínez und Scheffel 2012).
- Vorarbeiten: Was müssen Sie tun, bevor Sie annotieren können? Erfahren Sie, wie Sie ein CATMA-Projekt anlegen und Ihre Ressourcen im Tool verwalten.
- Funktionen: Welche Funktionen bieten Ihnen die CATMA-Projekte und das *Annotate*-Modul? Lernen Sie die einzelnen Komponenten des Tools kennen und lösen Sie Beispielaufgaben.
- Lösungen zu den Beispielaufgaben: Haben Sie die Beispielaufgaben richtig gelöst? Hier finden Sie Antworten.

1. Anwendungsbeispiel

In dieser Lerneinheit werden Sie in die grundlegenden Projektkoordinations- und Annotations-Funktionen des Tools CATMA (Schumacher 2024) eingeführt. CATMA (kurz für Computer Assisted Text Markup (vgl. Markup Language) and Analysis) ist ein webbasiertes (vgl. Webanwendung) und frei verfügbares (vgl. Open Access) Annotations-, Analyse- und Visualisierungstool für Texte. Das Tool eignet sich insbesondere für literaturwissenschaftliche Anwendungsfälle, kann aber auch für stärker formalisierte (bspw. linguistische) Annotationen verwendet werden. Unser Anwendungsbeispiel befasst sich mit einem erzähltheoretischen Phänomen: Anhand der späten Erzählung *Erstes Leid* (1924) von Franz Kafka werden Sie mit CATMA die Distanz der Erzählinstanz zum Erzählten anhand unterschiedlicher Rede- und Gedankenwiedergabeformen annotieren. Zugrunde liegt hierbei die Form der manuellen Annotation (Jacke 2024a), bei der Hervorhebungen oder Anmerkungen digital angebracht (und danach potentiell verfeinert und weiterverarbeitet) werden.

2. Vorarbeiten

Zunächst besorgen Sie sich die digitale Version von Kafkas Erzählung *Erstes Leid*, die wir in dieser Lerneinheit annotieren wollen. Folgen Sie dazu diesem Link zum TextGrid Repository (Horstmann 2024) und klicken in der linken Leiste in der Sektion „Herunterladen" auf „TEI-Corpus" (vgl. Korpus; XML) (siehe Abb. 1).

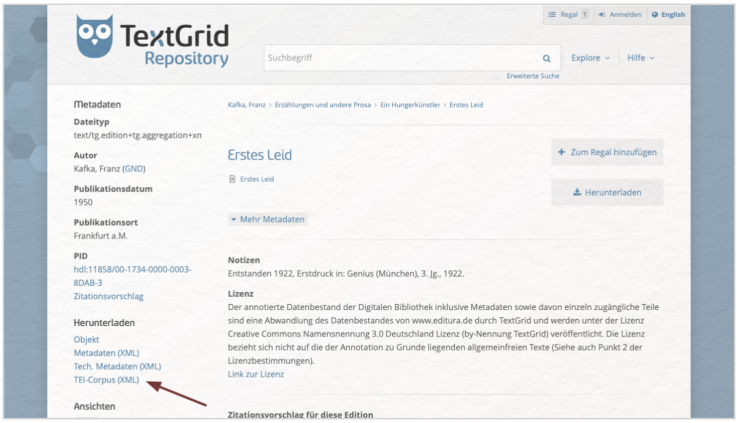

Abb. 1: Download des Textdigitalisats von Franz Kafkas „Erstes Leid“ im TextGrid Repository

Die heruntergeladene XML-Datei des Textes findet sich nun im Downloadordner Ihres Computers. Sie müssen diese nicht weiter bearbeiten, sondern können sie so, wie sie ist, später in CATMA hochladen.
Gehen Sie als nächstes auf https://catma.de, um sich im Tool anzumelden (siehe Abb. 2). Auf der Webseite können Sie sich bei Interesse über die einzelnen Funktionskomplexe des Tools, seine Entwicklungsgeschichte oder die zugrundeliegende Philosophie des „undogmatischen“ Annotierens informieren. Außerdem gibt es Tutorials auf Englisch, ein Manual und die Möglichkeit, den Newsletter zu abonnieren. Der Newsletter informiert etwa über Workshopangebote oder Neuerungen im Tool (so sind beispielsweise eine vereinfachte Kommentarfunktion, ein kategorieloses Annotieren und weitere Visualisierungsmöglichkeiten geplant).

Abb. 2: Die CATMA-Webseite

CATMA ist webbasiert und muss daher nicht lokal installiert werden. Sie benötigen lediglich eine stabile Internetverbindung. Klicken Sie jetzt auf „Work with CATMA 6“, um zu dieser Ansicht (vgl. GUI) zu gelangen:

Abb. 3: Die Sign-in-Seite von CATMA

Sie haben nun zwei Möglichkeiten. Wenn Sie eine Googlemailadresse besitzen, können Sie diese zur **Authentifizierung** benutzen und direkt auf den Button „Sign in“ klicken und dann – nach Bestätigung der „terms of use“, die Sie dort verlinkt finden – auf „LOG IN WITH YOUR GOOGLE ACCOUNT“. CATMA nutzt lediglich

Googles Authentifizierungsmechanismus und sonst nichts: Die von Ihnen hochgeladenen und im Tool erstellten Daten werden auf einen sicheren Server der Universität Hamburg geladen und können von keiner dritten Partei eingesehen werden. Sollten Sie lieber einen eigenen CATMA-Account erstellen wollen, klicken Sie zunächst auf den „Sign up"-Button und geben eine gültige Emailadresse ein. An diese Adresse erhalten Sie umgehend einen Aktivierungscode. Nach dem Klick auf den Link in der Email können Sie einen selbst gewählten Usernamen und ein Passwort eingeben. Mit diesen Daten können Sie sich in Zukunft durch Klick auf den „Sign in"-Button (siehe Abb. 3) bei CATMA anmelden.
Nach dem Login landen Sie im **Home-Bereich** von CATMA (siehe Abb. 4). Hier können Sie neue Projekte anlegen oder – insofern Sie kollaborativ arbeiten, was wir in dieser Lerneinheit *nicht* tun – bestehenden Projekten beitreten. Oben rechts sollten Sie einmal auf „Edit Account" klicken und ggf. Ihren Usernamen anpassen: Da CATMA die kollaborative Arbeit an Projekten ermöglicht, kann Ihr Username beim Teilen von Ressourcen als Suchvorschlag erscheinen. Um das Teilen von Ressourcen zu vereinfachen, empfiehlt es sich, Ihren Realnamen auch als Usernamen zu verwenden; prinzipiell können Sie hier jedoch selbst entscheiden. In jedem Fall sollte Ihr Username keine sensiblen Daten enthalten.

Abb. 4: CATMA-Startseite: Organisation von Projekten

Als nächstes werden Sie ein **Projekt** für die Annotation von Kafkas Erzählung **anlegen**, indem Sie auf „CREATE NEW PROJECT" klicken. Wenn Sie einen Projektnamen (wir haben uns hier für „forTEXT-Lerneinheit" entschieden) und ggf. eine Beschreibung eingegeben haben, erscheint Ihr neu angelegtes Projekt auf der Startseite (siehe Abb. 5). Um zu einem späteren Zeitpunkt auf diese Startseite zurückzukehren und Ihre Projekte zu organisieren, klicken Sie einfach auf das oben links erscheinende Home-Symbol.

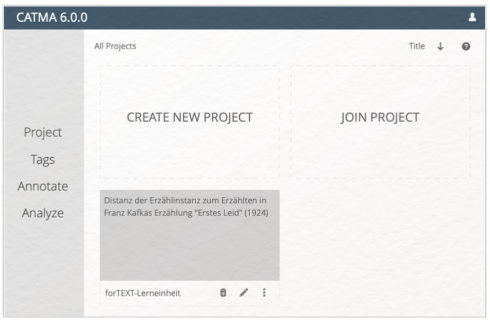

Abb. 5: Die CATMA-Startseite mit neu angelegtem Projekt

Klicken Sie anschließend auf die Kachel, die Ihr neues Projekt anzeigt, betreten Sie das erste **Modul** in CATMA: „***Project***". In welchem Modul Sie sich befinden, wird stets in der linken grauen Spalte angezeigt. Das *Project*-Modul zeigt Ihnen, welche Ressourcen in Ihrem Projekt enthalten sind, wer die Projektmitglieder sind und welche Rollen diese Mitglieder haben (siehe Abb. 6). Da wir im aktuellen Anwendungsbeispiel nicht kollaborativ arbeiten, werden in diesem Fall lediglich Sie als „Owner" des Projektes angezeigt.

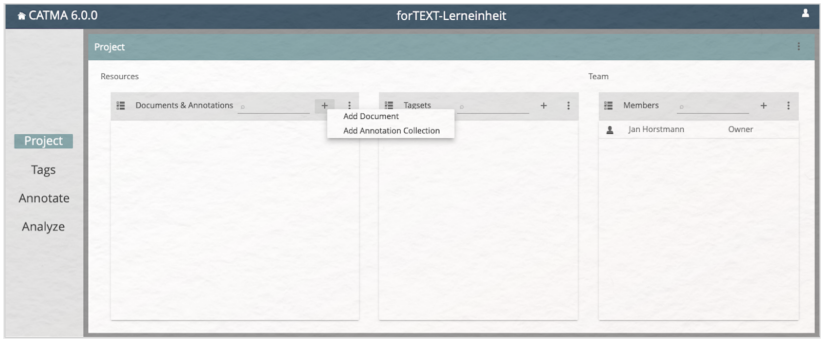

Abb. 6: Das Project*-Modul in CATMA*

Mit einem Klick auf das „+“-Symbol neben „Documents & Annotations“ haben Sie die Möglichkeit, Ihrem Projekt die Erzählung *Erstes Leid* hinzuzufügen. Klicken Sie dazu auf „**Add Document**“.

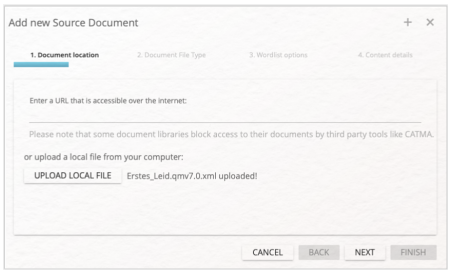

Abb. 7: Dokumentupload in CATMA

Es öffnet sich ein Fenster, in dem Sie auf „UPLOAD LOCAL FILE“ klicken und in der Ordnerstruktur Ihres Computers nach der zuvor heruntergeladenen XML-Datei von Kafkas Erzählung suchen. CATMA verarbeitet alle gängigen Textdateiformate. Wenn Sie mehrere Dokumente auf einmal hochladen möchten, fassen Sie diese Dokumente zunächst in einem ZIP-Ordner (vgl. ZIP) zusammen. CATMA wird den Ordner im Upload-Prozess wieder entpacken und jedes enthaltene Dokument einzeln anzeigen. Klicken Sie danach auf den „NEXT“-Button unten rechts. Die nächste Ansicht zeigt Ihnen dann eine Vorschau des hochzuladenden Textes (siehe Abb. 8). Der „File Type“ und das „Encoding“ in dieser Ansicht werden i. d. R. automatisch korrekt angezeigt. Sollte sich Veränderungsbedarf ergeben (was in dieser Lerneinheit nicht der Fall ist), können Sie diese beiden Felder durch Doppelklick bearbeiten.

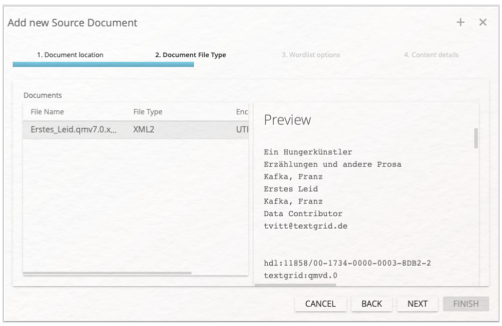

Abb. 8: Preview des hochzuladenden Dokuments in CATMA

Nach einem Klick auf „NEXT“ gelangen Sie zu den Einstellungsmöglichkeiten für Sprache, die ebenfalls im Regelfall automatisch korrekt angezeigt werden sollte (siehe Abb. 9). Erfahrenere Nutzer*innen können unter „Advanced Options“ noch individuell bestimmen, welche Buchstabenabfolgen (wie etwa „z. B.“) vom System als einzelnes Wort behandelt werden sollen. Diese Option kann für die Lerneinheit vernachlässigt werden. Klicken Sie wieder auf „NEXT“.

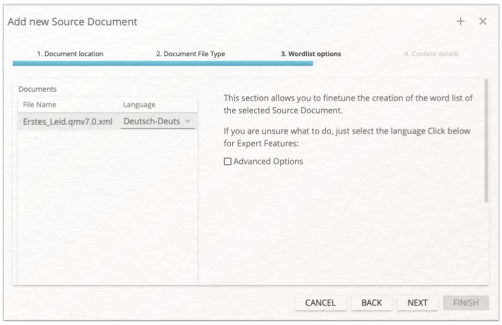

Abb. 9: Spracheinstellungen im Dokument-Upload von CATMA

Im letzten Fenster des Upload-Prozesses (siehe Abb. 10) haben Sie die Möglichkeit, Details zu dem hochgeladenen Dokument anzugeben. Dies ist insbesondere sinnvoll, wenn Sie mit mehreren Dokumenten arbeiten, die ggf. aus

unterschiedlichen Editionen stammen. Das Fenster ermöglicht ihnen, auch später einen Überblick zu halten. Die einzelnen Felder können Sie per Doppelklick bearbeiten (klicken Sie danach auf „Save“).

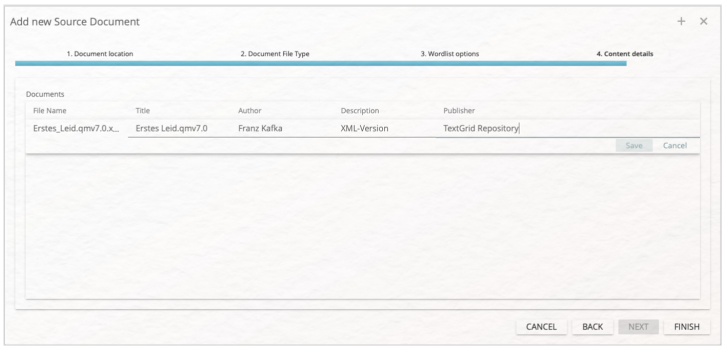

Abb. 10: Dokument-Details bearbeiten im Upload-Prozess in CATMA

Ein Klick auf den „FINISH“-Button rechts unten beendet den Upload-Prozess des Dokuments und Sie landen wieder im *Project*-Modul. Sie haben nun erfolgreich einen CATMA-Account und ein CATMA-Projekt angelegt sowie Ihren ersten Text hochgeladen.

3. Funktionen

Um annotieren zu können, fehlt Ihnen jetzt noch die Taxonomie der Kategorien, anhand derer der Text untersucht und annotiert werden soll. Solche Taxonomien heißen in CATMA **Tagsets** (vgl. Tagset). Sie erstellen ein Tagset, indem Sie auf das „+“-Zeichen klicken und einen Namen für das zu erstellende Tagset vergeben (wir haben uns hier für „Distanz (Erzählinstanz–Erzähltes)“ entschieden; siehe Abb. 11). Sie können Tagsets (wie auch Documents, Annotation Collections oder Members) jederzeit über die drei kleinen Punkte neben den „+“-Zeichen im *Project*-Modul bearbeiten.

Abb. 11: Erstellen eines Tagsets in CATMA

Ein Doppelklick auf das erstellte Tagset in der entsprechenden Spalte im *Project*-Modul oder auch ein Klick auf „Tags“ in der linken Navigationsspalte erlaubt Ihnen, das Tagset im ***Tags*-Modul** mit einzelnen Tags (d. h. den Annotationskategorien) auszustatten (siehe Abb. 12). Übrigens: Für die Navigation in CATMA brauchen Sie niemals die Vor- und Zurück-Buttons Ihres Browsers (vgl. Browser) zu bedienen, sondern sollten immer die entsprechenden Schaltflächen auf der linken Seite des CATMA-Fensters nutzen.

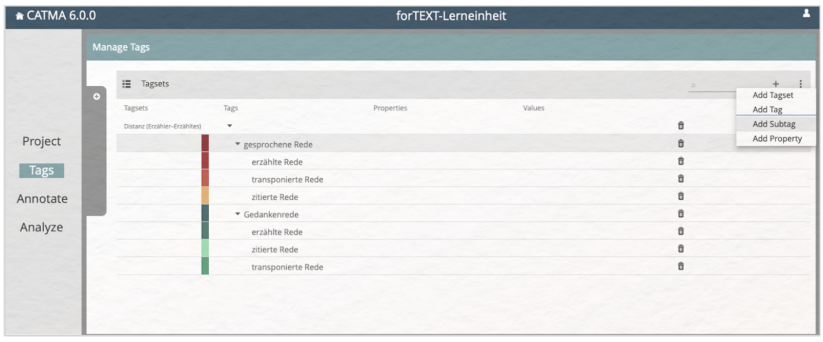

Abb. 12: Das Tags*-Modul in CATMA*

Im *Tags*-Modul können Sie das eben erstellte Tagset auswählen und beliebig viele Kategorien hinzufügen; CATMA macht diesbezüglich keine Einschränkungen oder Vorschriften (auch deswegen nennt es sich „undogmatisch“). Für diese Lerneinheit haben wir uns für die von Martínez und Scheffel (2012, 65) in ihrer *Einführung in die Erzähltheorie* definierten Kategorien zur Analyse der sog. „Distanz“ entschieden, die im Wesentlichen auf unterschiedliche Rede- bzw. Gedankenwiedergabeformen zurückgeht. Die Leitfrage lautet dabei: Wie nah ist die Erzählinstanz sprachlich, gedanklich, ideologisch an der von ihr vermittelten ‚Welt‘ und ihren Figuren positioniert? Selbstverständlich könnte man die Erzählung auch nach allen möglichen anderen Kategorien untersuchen, wie beispielsweise Kategorien der künstlerischen Existenz und des Verständnisses von Kunst (Vollmer 1998), Kategorien des unzuverlässigen Erzählens (Gaasland 2012) oder auch Emotionskategorien, um die Erzählung einer Sentimentanalyse (Flüh 2024) im Sinne eines Close Readings (vgl. Close Reading) zu unterziehen etc. Für viele dieser Ansätze ließen sich analog zu den hier ausgewählten Distanzkategorien mehr oder weniger hierarchisch organisierte Taxonomien in Form von Tagsets entwickeln. Sowohl in Bezug auf die *gesprochene Rede* als auch auf die *Gedankenrede* unterscheiden Martínez und Scheffel (2012) drei Kategorien: die *erzählte*, die *transponierte* und die *zitierte Rede*. Diese drei Unterkategorien haben noch weitere Unterkategorien, die jedoch für die hier zu vollziehende Übung nicht weiter ausdifferenziert werden müssen. Die weiteren Unterkategorien sind „Erwähnung des sprachlichen Akts“, „Gesprächsbericht“ bzw. „Bewusstseinsbericht“ für die Kategorie der erzählten Rede, „indirekte Rede“, „erlebte Rede“ für die Kategorie der transponierten Rede und „direkte Rede“, „autonome direkte Rede“ bzw. „Gedankenzitat“ und „innerer Monolog“ für die Kategorie der zitierten Rede; vgl. Martínez und Scheffel (2012, 65). Aus diesen Kategorien lässt sich nun leicht ein Tagset entwerfen, wie es in Abb. 12 beispielhaft geschehen ist. Wählen Sie das Tagset aus, klicken oben rechts auf das „+“-Symbol und dann auf „Add Tag“. So erstellen Sie die beiden Tags „gesprochene Rede“ und „Gedankenrede“ (siehe Abb. 13). Um die jeweiligen Unterkategorien zu erstellen, wählen Sie den entsprechenden Tag aus und klicken dann oben rechts auf „Add Subtag“ und die neu erstellten Kategorien erscheinen eine Hierarchieebene weiter eingerückt.

Abb. 13: Erstellen eines Tags in CATMA

Sie können potentiell auf diese Art und Weise so viele Tags auf so vielen Ebenen erstellen wie Sie möchten – nur die Übersichtlichkeit und praktische Handhabbarkeit setzen Ihnen hier Grenzen. Auch die Farben der einzelnen Kategorien können Sie frei wählen. Und: Tagsets und einzelne Tags können später im Annotationsprozess jederzeit editiert und erweitert werden.

Aufgabe 1: Bauen Sie das Tagset wie in Abb. 12 angezeigt nach. Was könnten Ihrer Meinung nach die Vor- und Nachteile bzw. Gefahren von Konzeptrepräsentationen im Sinne formalisierter Taxonomien in Form von Tagsets sein?

In CATMA ist die Unterscheidung zwischen **Tags** und **Annotationen** grundlegend. Ähnlich wie in der linguistischen Unterscheidung von **Tagsets** bezeichnen Tags jeweils die generelle Kategorie und Annotationen die spezifischen Vorkommnisse dieser Kategorie im Text. Neben den text-unspezifischen Taxonomien, wie sie durch Tagsets abgebildet werden, wollen wir jetzt textspezifische Annotationen mithilfe dieser Kategorien erstellen.
Eine technische Besonderheit von CATMA ist, dass Annotationen nicht direkt im Textdokument gespeichert werden (sog. *inline Markup*), sondern in einer mit dem Originaltextdokument verknüpften Datenbank abgelegt werden (sog. *external stand-off Markup*). Diese Datenbank erfasst alle Annotationen aller Nutzer*innen, die sich auf das jeweilige Dokument beziehen. So kann erstens ein und derselbe Text von beliebig vielen Annotator*innen und mit beliebig vielen Taxonomien ausgezeichnet werden. Diese kollaborativ erstellte Gesamtmenge aller Annotationen kann man dann zweitens bei späteren Suchabfragen (vgl. Query) nutzen. Um beides technisch zu ermöglichen, werden Annotationen in einer sog. ***Annotation Collection*** gespeichert, die jeweils einem spezifischen Text und einer Person zugeordnet ist.
Es gibt in CATMA mehrere Möglichkeiten, Annotation Collections zu erstellen: Sie können beispielsweise zum *Project*-Modul zurückkehren und in der „Documents“-Spalte Ihrem Dokument eine Annotation Collection hinzufügen. Das funktioniert aber auch während des eigentlichen Annotationsprozesses, wie wir es auch in dieser Lerneinheit tun. Klicken Sie daher nun in der linken Navigationsspalte auf „Annotate“, um das ***Annotate*-Modul** (siehe Abb. 14) zu betreten.

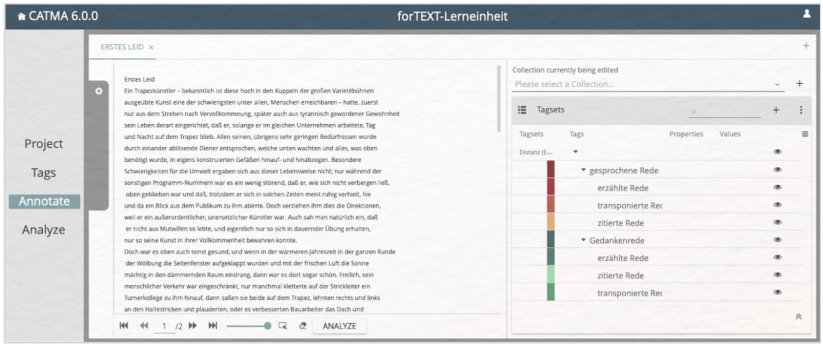

Abb. 14: Das Annotate*-Modul in CATMA mit geöffnetem Tagset*

Oben rechts im „Collection"-Feld haben Sie die Möglichkeit, mit einem Klick auf das „+"-Symbol Ihre eigene Annotation Collection zu erstellen (siehe Abb. 15). Da Annotation Collections sowohl text- als auch personenspezifisch sind, empfiehlt es sich, diesen Bezug bei der Benennung Ihrer Annotation Collection explizit zu machen.

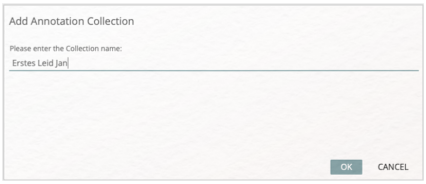

Abb. 15: Erstellen einer Annotation Collection in CATMA

Nach einem Klick auf „OK" haben Sie alles, was benötigt wird, und können nun mit dem eigentlichen **Annotieren** beginnen.

Um ein Wort oder eine Textpassage zu annotieren, markieren Sie zunächst den entsprechenden Abschnitt, sodass er blau hinterlegt ist. Jetzt gibt es zwei Möglichkeiten:

1. Klicken Sie auf die gewünschte Kategorie im Tagset auf der rechten Seite. Die Annotation wird in der Farbe des ausgewählten Tags als Unterstreichung erscheinen (siehe Abb. 16).
2. Alternativ können Sie auch mit rechts in das Textfeld klicken. Es öffnet sich dort ein komprimiertes Menü Ihres Tagsets, aus dem Sie den entsprechenden Tag auswählen können (siehe Abb. 17).

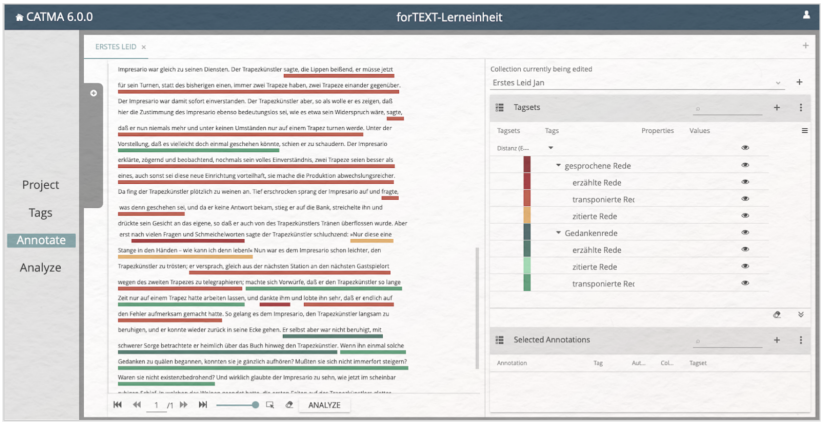

Abb. 16: Das Annotate-Modul in CATMA mit erstellten Annotationen

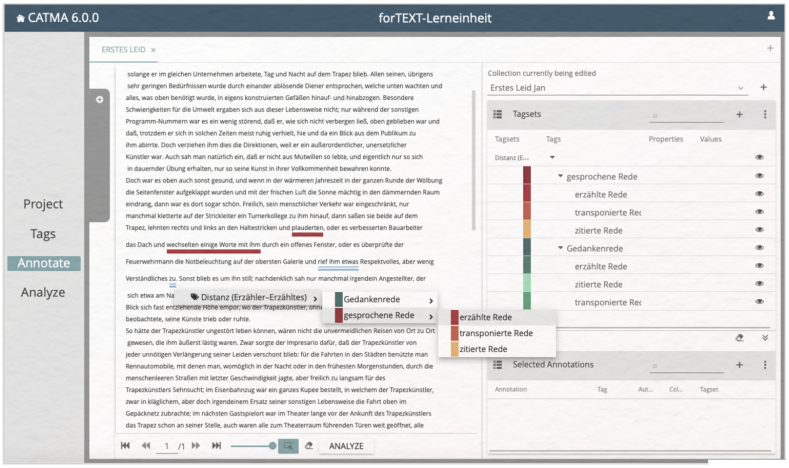

Abb. 17: Erstellen einer Annotation mit Rechtsklick in CATMA

Es gibt keine Begrenzung, wie viele Annotationen Sie vergeben können oder wie lang eine Annotation sein soll. Auch können Sie ein und dieselbe Textstelle mit unterschiedlichen Kategorien belegen und Annotationen können einander auch überlappen. Wenn sie voneinander getrennte Passagen mit einer einzigen Annotation belegen wollen (wie etwa „rief ihm etwas […] zu“), gibt es die Möglichkeit, unterhalb des Textfeldes die Option „Allow multiple discontinuous selections“ (rechts neben dem Zoom-Slider; siehe Abb. 17) zu aktivieren. Der Slider bestimmt die Seitengröße, so ist es beispielsweise möglich, entweder den gesamten Text auf einer Seite scrollbar anzeigen zu lassen (Slider auf 100) bzw. einzelne Seiten zu erzeugen, die Sie dann mit den entsprechenden Schaltflächen durchblättern können.
Haben Sie eine **Annotation** irrtümlich gesetzt und möchten sie wieder **löschen**, klicken Sie einfach die entsprechende Annotation im Textfeld an. Unten rechts erscheint ein Fenster mit „Selected Annotations“, in dem Sie die Möglichkeit haben, mit Klick auf das kleine Mülltonnen-Symbol die ausgewählte Annotation wieder zu löschen.

Aufgabe 2: Lesen Sie die Erzählung *Erstes Leid* am Bildschirm und annotieren Sie Passagen, die zu den gegebenen Distanz-Kategorien passen. Was ist in Bezug auf die vergebenen Kategorien auffällig? Was sind Vor- und Nachteile eines solchen digital gestützten Annotationsprozesses?

Sie werden im Annotationsprozess bemerkt haben, dass Sie sich in einigen Fällen sehr sicher sind, wie etwas annotiert werden muss, in anderen aber zumindest Diskussionsbedarf bestünde (Ist z. B. die Erwähnung, dass der Trapezkünstler träumt, bereits ein Gedankenbericht?). Um derlei Abwägungen ebenfalls in den Metadaten, d. h. den Annotationen sichtbar machen zu können, bietet CATMA die Möglichkeit, sog. **Tagsets** (vgl. Tagset) und ggf. ***Values*** zu vergeben. Properties (= Eigenschaften) ermöglichen, das tendenziell deklarativ organisierte Annotieren mit Tags um ein skalares Konzept qualitativer Bewertungen zu erweitern, d. h. um Kategorien, die auf unterschiedlichsten Ebenen des Tagsets auftauchen können. So ließen sich beispielsweise jedem Tag die Properties „**Sicherheit**“ oder auch „**Wichtigkeit**“ etc. hinzufügen. Ist eine solche Property einem Tag zugeordnet, wird das System Sie nach jeder Vergabe des entsprechenden Tags nach dieser Property, also nach der Sicherheit, Wichtigkeit etc. der gesetzten Annotation, fragen. Hier könnten Sie jetzt sog. „ad hoc values“ vergeben, die näher beschreiben, wie sicher eine Annotationsentscheidung war. Es ist auch möglich, bereits vorab zu bestimmen, welche Values bei einer bestimmten Property ausgewählt werden können. Im Falle der „Sicherheits“-Property könnten Sie beispielsweise eine Skala von 1 bis 5 vorgeben, sodass die Sicherheiten der einzelnen Annotationen am Ende tatsächlich vergleichbar und damit messbar gemacht werden. Dies alles liegt in Ihrem eigenen Ermessen und orientiert sich an den Bedarfen Ihres jeweiligen Annotationsprojektes.
Properties können Sie ebenfalls im *Tags*-Modul vergeben. Wählen Sie dazu einen Tag aus und klicken oben beim „+“-Symbol auf „Add Property“. Im sich öffnenden Fenster vergeben Sie einen Property-Namen, klicken auf die Schaltfläche „ADD PROPERTY“ und vergeben dann ggf. dazugehörige (durch Kommata separierte) Values (siehe Abb. 18). Mit einem Klick auf „OK“ werden die Properties und Values erstellt.

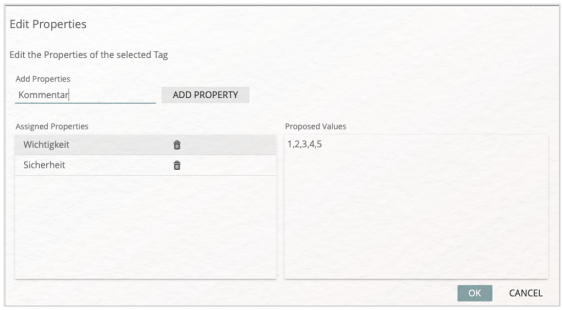

Abb. 18: Hinzufügen einer Property *und entsprechenden* Values *zu einem Tag in CATMA*

Es ist zudem möglich, die Property-Funktion für **Freitextkommentare** zu nutzen. Vergeben Sie dafür einfach eine Property mit dem Namen „Kommentar" für diejenigen Tags, bei denen Sie Kommentarbedarf sehen. Bei Vergabe einer entsprechenden Annotation wird künftig nach dieser Property gefragt und Sie haben die Möglichkeit, im Value-Feld einen annotationsspezifischen Kommentar zu hinterlassen, der bspw. Ihre Annotationsentscheidung näher erläutert oder rechtfertigt. Derlei Kommentare können auch für einen selbst eine wertvolle Gedächtnisstütze für spätere Annotations- oder Analysedurchgänge bilden.

Aufgabe 3: Gehen Sie Ihre Annotationen noch einmal durch und vergeben Sie Properties und Values. Sie können die in dieser Lerneinheit bereits vorgeschlagenen Kategorien („Sicherheit", „Wichtigkeit", „Kommentar") nutzen, oder auch jede weitere Form von Property erstellen, die Ihnen wichtig erscheint. Welche Textstellen bedürfen für die Annotation der weiteren Diskussion? Welche weiteren Properties bieten sich an?

Tipp: Wenn Sie in einem CATMA-Projekt mit mehreren Texten, mehreren Tagsets oder auch mehreren Annotation Collections arbeiten (Letzteres ist insbesondere bei der kollaborativen Annotation (Jacke 2024b) der Fall, die in dieser Lerneinheit ausgespart wurde), hilft Ihnen die Lasche im *Annotate*-Modul, einen Übersicht zu behalten (siehe Abb. 19). Klicken Sie links auf die Lasche, um sie zu öffnen, wählen Sie die gewünschten Texte, Annotation Collections und Tagsets aus und schließen die Lasche wieder.

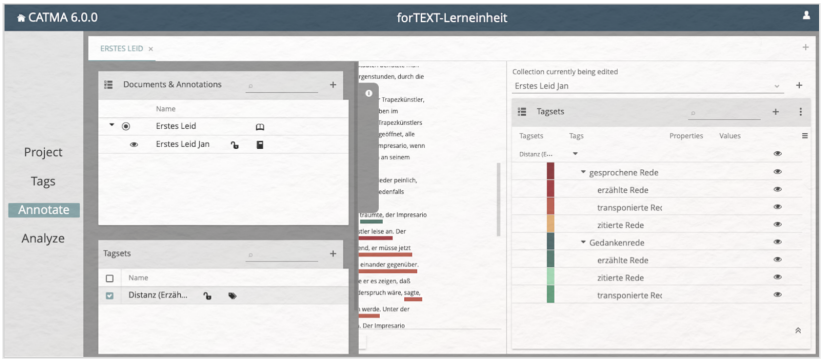

Abb. 19: Die Lasche in CATMAs Annotate*-Modul zur Organisation von Dokumenten, Annotation Collections und Tagsets*

Wie Sie die von Ihnen dem Text als Metadaten hinzugefügten Annotationen zusammen mit den Textdaten analysieren und visualisieren können, werden wir Ihnen in der nächsten Lerneinheit erklären.

4. Lösungen zu den Beispielaufgaben

Aufgabe 1: Bauen Sie das Tagset, wie in Abb. 12 angezeigt, nach. Was könnten Ihrer Meinung nach die Vor- und Nachteile bzw. Gefahren von Konzeptrepräsentationen im Sinne formalisierter Taxonomien in Form von Tagsets sein?

Entwickeln Sie selbst basierend auf einem konkreten Analyseansatz oder einer Theorie ein Tagset, werden Sie merken, dass man schnell dazu neigt, zu viele Kategorien anzulegen. Es kann dann schwierig werden, beim eigentlichen Annotationsprozess den Überblick über das eigene Tagset und sämtliche zur Verfügung stehenden Tags im Kopf zu behalten. Häufig macht die Formalisierung von Konzepten in Form von Taxonomien aber auch recht unmittelbar auf Doppelungen, Lücken oder Ungenauigkeiten in der eigenen Taxonomie aufmerksam. Nicht selten kann man mehrere Tags einer Oberkategorie hinzufügen, wodurch sie an anderer Stelle vielleicht obsolet werden, sodass eine Reduktion des Tagsets nicht nur die Übersichtlichkeit wieder herstellt, sondern

auch für eine präzisere Organisation der gesamten Taxonomie sorgt.
Ihnen wird auch auffallen, dass die Erzeugung von Tagkategorien vor dem Hintergrund der konkreten Anwendung an Textpassagen bewusst macht, dass jeder Kategorie sehr genaue Definitionen zugrunde liegen sollten, damit – auch Ihnen selbst – klar wird, wie und wann der jeweilige Tag anzuwenden ist. Die Hierarchisierung einzelner Tags passt zudem zu einigen literaturwissenschaftlichen Ansätzen besser (wie etwa dem Formalismus oder Strukturalismus) als zu anderen (wie etwa der Dekonstruktion). Dass Tagsets in CATMA nicht notwendigerweise hierarchisch organisiert werden müssen, schafft diesem Problem Abhilfe. Auf derselben Hierarchieebene angesiedelte Tags werden in CATMA zwar immer noch horizontal untereinander angezeigt, dennoch können sie beispielsweise eine rhizomatische Konzeptorganisation simulieren. Eine allgemeine Gefahr der Formalisierung von Konzepten in Form von Tagsets könnte sein, dass (möglicherweise bewusste) Uneindeutigkeiten in der eigenen Taxonomie schwerer darstellbar sind, da die tabellarische Darstellung zu einer Verdeutlichung zwingt, die gelegentlich mit einer Simplifizierung des Analysekonzeptes selbst einhergehen könnte.

Aufgabe 2: Lesen Sie die Erzählung *Erstes Leid* am Bildschirm und annotieren Sie Passagen, die zu den gegebenen Distanz-Kategorien passen. Was ist in Bezug auf die vergebenen Kategorien auffällig? Was sind Vor- und Nachteile eines solchen digital gestützten Annotationsprozesses?

Es ist auffällig, dass zu Beginn des Textes keine der gegebenen Distanz-Kategorien angewendet werden kann. Schließlich finden sich Gesprächsberichte (erzählte Rede), die durch indirekte und erlebte Rede abgelöst werden (geringere Distanz). Direkte Rede (Martínez und Scheffel 2012) („dramatischer Modus“) findet sich nur ein einziges Mal, wenn das Leid des Trapezkünstlers am größten ist: „Nur diese eine Stange in den Händen – wie kann ich denn leben!“ (übrigens gekoppelt mit dem einzigen Ausrufungszeichen des Textes). Anschließend folgen indirekte Gedankenwiedergaben und Bewusstseinsberichte (des Impresarios) – die Distanz wird also wieder größer. Die Distanz-Entwicklung des gesamten Textes ließe sich folglich als Zoom-in und anschließenden Zoom-out beschreiben, wobei im Moment der geringsten Distanz der Ursprung des „Erste[n] Leid[s]“ liegt. Der literarische Text entzieht sich gelegentlich den gegebenen Analysekategorien. An diesen Stellen ist nicht klar, ob beispielsweise Gedanken repräsentiert werden oder ob Werturteile auf Gedanken (wessen?) bzw. Gesprächen basieren. Je direkter die Form der Gedanken-/Redewiedergabe ist, desto unproblematischer lassen sich die Tags vergeben. Wird die Distanz der Erzählinstanz zum Erzählten jedoch größer (insbesondere bei der bloßen Erwähnung eines sprachlichen Aktes oder Gedankens), kann es zu Entscheidungsschwierigkeiten kommen.
Man könnte weitere Kategorien und Unterkategorien (sog. Subtags) vergeben, um eine genauere Annotation zu ermöglichen. Die Gefahr dabei ist, ein unübersichtlich großes Tagset zu erstellen, dessen analytische Aussagekraft immer in seiner Differenzierbarkeit begründet liegt. Der große Vorteil ist, dass der Prozess systematisiert wird. Anmerkungen von potentiell mehreren Benutzer*innen werden mit dem Text verknüpft und so nachhaltig und analysierbar gemacht.

Aufgabe 3: Gehen Sie Ihre Annotationen noch einmal durch und vergeben Sie Properties und Values. Sie können die in dieser Lerneinheit bereits vorgeschlagenen Kategorien („Sicherheit“, „Wichtigkeit“, „Kommentar“) nutzen, oder auch jede weitere Form von Property erstellen, die Ihnen wichtig erscheint. Welche Textstellen bedürfen für die Annotation der weiteren Diskussion? Welche weiteren Properties bieten sich an?

Textstellen zur weiteren Diskussion in Bezug auf Distanzkategorien könnten sein:

1. „und wenn in der wärmeren Jahreszeit in der ganzen Runde der Wölbung die Seitenfenster aufgeklappt wurden und mit der frischen Luft die Sonne mächtig in den dämmernden Raum eindrang, dann war es dort sogar schön“
2. „der Trapezkünstler lag im Gepäcknetz und träumte“
3. „Da fing der Trapezkünstler plötzlich zu weinen an“
4. „da er keine Antwort bekam“

zu 1) Auf wessen Wahrnehmung bzw. Gedanken beruht das Urteil „schön“? Wird hier bereits auf einen Gedankengang des Trapezkünstlers verwiesen, sodass von einem Bewusstseinsbericht die Rede sein könnte? Auch innerhalb der Unterkategorie Bewusstseinsbericht lassen sich demnach noch unterschiedliche Distanzen ausmachen. Entscheiden Sie sich dazu, diese Passage mit dem Gedankenrede-Tag „erzählte Rede“ zu annotieren, sollten Sie diese Abwägungen durch einen niedrigen Wert des „Sicherheits“-Propertys verdeutlichen.

zu 2) Ähnlich wie bei (1) ist hier nicht klar, ob mit der Erwähnung des Träumens Bewusstsein berichtet wird: Diskussionen über Traumtheorien, das Bewusste, Unbewusste, Halb-Bewusste, Unterbewusste etc. könnten sich anschließen. Auch hier könnte die Unsicherheit durch einen niedrigen Wert des „Sicherheit“-Propertys verdeutlicht werden.

zu 3) Das Weinen als Reaktion setzt ein Bewusstsein voraus, auf das hier indirekt Bezug genommen wird. Wenn Sie sich entschließen, den entsprechenden Tag zu vergeben, bietet es sich an, Ihre Entscheidung in einem Kommentar zu dokumentieren.

zu 4) Wie geht man mit Textpassagen um, die das Nichtvorhandensein von gesprochener Rede oder Gedankenrede explizit hervorheben? Für die angeführte Stelle könnte man einen weiteren Tag erstellen („Schweigen“ o. ä.) und sie entsprechend annotieren. Die Schwierigkeit weist jedoch auf ein größeres Problem hin, das in den Digital Humanities bislang nicht geklärt wurde, für die Literaturwissenschaften aber ganz entscheidend ist: Wie kann im Text etwas annotiert werden, das nicht da ist?

Weitere Properties könnten etwa sein, wer in einem Gespräch die Gesprächspartner sind oder wessen Gedanken wiedergegeben werden.

Externe und weiterführende Links

- CATMA: https://web.archive.org/save/https://catma.de (Letzter Zugriff: 03.07.2024)
- Digitale Version von Kafkas Erzählung *Erstes Leid* auf TextGrid: https://web.archive.org/save/https://textgridrep.org/browse/-/browse/qmv7_0 (Letzter Zugriff: 03.07.2024)

Bibliographie

Flüh, Marie. 2024. Methodenbeitrag: Sentimentanalyse. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 7. Sentimentanalyse (7. Oktober). doi: 10.48694/fortext.3797, https://fortext.net/routinen/methoden/sentimentanalyse.

Gaasland, Rolf. 2012. Practical Reasoning Demarcated. Unreliable Narration in Franz Kafka’s „Erstes Leid“. In: *Disputable Core Concepts of Narrative Theory*, hg. von Göran Rossholm und Christer Johansson, 239–251. Bern (u.a.): Lang.

Horstmann, Jan. 2024. Ressourcenbeitrag: TextGrid Repository. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 11. Bibliografie (29. November). doi: 10.48694/fortext.3794, https://fortext.net/ressourcen/textsammlungen/textgrid-repository.

Jacke, Janina. 2024b. Methodenbeitrag: Kollaboratives literaturwissenschaftliches Annotieren. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3749, https://fortext.net/routinen/methoden/kollaboratives-literaturwissenschaftliches-annotieren.

———. 2024a. Methodenbeitrag: Manuelle Annotation. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3748, https://fortext.net/routinen/methoden/manuelle-annotation.

Martínez, Matías und Michael Scheffel. 2012. *Einführung in die Erzähltheorie*. München: Beck.

Schumacher, Mareike. 2024. Toolbeitrag: CATMA. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 4. Manuelle Annotation (7. August). doi: 10.48694/fortext.3761, https://fortext.net/tools/tools/catma.

Vollmer, Hartmut. 1998. Die Verzweiflung des Artisten. Franz Kafkas Erzählung Erstes Leid – eine Parabel künstlerischer Grenzerfahrungen. *Deutsche Vierteljahrsschrift für Literaturwissenschaft und Geistesgeschichte* 72, Nr. 1: 126–146. doi: 10.1007/BF03375489,.

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

Close Reading Close Reading bezeichnet die sorgfältige Lektüre und Interpretation eines einzelnen oder weniger Texte. Close Reading ist in der digitalen Literaturwissenschaft außerdem mit der manuellen Annotation textueller Phänomene verbunden (vgl. auch Distant Reading als Gegenbegriff).

Commandline Die Commandline (engl. *command line interface* (CLI)), auch Kommandozeile, Konsole, Terminal oder Eingabeaufforderung genannt, ist die direkteste Methode zur Interaktion eines Menschen mit einem Computer. Programme ohne eine grafische Benutzeroberfläche (GUI) werden i. d. R. durch Texteingabe in die Commandline gesteuert. Um die Commandline zu öffnen, klicken Sie auf Ihrem Mac „cmd“ + „space“, geben „Terminal“ ein und doppelklicken auf das Suchergebnis. Bei Windows klicken Sie die Windowstaste + „R“, geben „cmd.exe“ ein und klicken Enter.

Distant Reading Distant Reading ist ein Ansatz aus den digitalen Literaturwissenschaften, bei dem computationelle Verfahren auf häufig große Mengen an Textdaten angewandt werden, ohne dass die Texte selber gelesen werden. Meist stehen hier quantitative Analysen im Vordergrund, es lassen sich jedoch

auch qualitative Metadaten quantitativ vergleichen. Als Gegenbegriff zu *Close Reading* wurde der Begriff insbesondere von Franco Moretti (2000) geprägt.

Feature Unter Features können Einzelfunktionen eines Tools verstanden werden, die beispielsweise komplexe Funktionen wie die Visualisierung eines Textes als Wordcloud ermöglichen, oder auch kleinere Funktionseinheiten wie den Abgleich einzelner Spracheigenschaften (Properties) mit annotierten Beispieltexten darstellen.

GUI GUI steht für *Graphical User Interface* und bezeichnet eine grafische Benutzeroberfläche. Ein GUI ermöglicht es, Tools mithilfe von grafischen Schaltflächen zu bedienen, um somit beispielsweise den Umgang mit der Commandline zu umgehen.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

Korpus Ein Textkorpus ist eine Sammlung von Texten. Korpora (Plural für „das Korpus“) sind typischerweise nach Textsorte, Epoche, Sprache oder Autor*in zusammengestellt.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup (Textauszeichung) Die Textauszeichnung (eng. *Markup*) fällt in den Bereich der Daten- bzw. Textverarbeitung, genauer in das Gebiet der Textformatierung, welche durch Auszeichnungssprachen wie XML implementiert wird. Dabei geht es um die Beschreibung, wie einzelne Elemente eines Textes beispielsweise auf Webseiten grafisch dargestellt werden sollen.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z.B. HTML, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch Annotationen durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch Annotationen bzw. Markup sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

Open Access Open Access bezeichnet den freien Zugang zu wissenschaftlicher Literatur und anderen Materialien im Internet.

POS PoS steht für *Part of Speech* , oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

Property Property steht für „Eigenschaft“, „Komponente“ oder „Attribut“. In der automatischen Annotation dienen konkrete Worteigenschaften wie Groß- und Kleinschreibung zur Klassifizierung von Wörtern oder Phrasen. Durch die Berücksichtigung solcher Eigenschaften in den Features eines Tools kann maschinelles Lernen bestimmter Phänomene umgesetzt werden. In der manuellen Annotation können als Properties auch Eigenschaften von Annotationen benannt werden.

Query *Query* bedeutet „Abfrage“ oder „Frage“ und bezeichnet eine computergestützte Abfrage zur Analyse eines Textes. Um Datenbestände zu durchsuchen, werden Abfragesprachen eingesetzt, die *Queries* (Anfragen) an den Datenbestand senden. So bilden alle möglichen Queries zusammen die *Query Language* eines Tools.

Server Ein Server kann sowohl hard- als auch softwarebasiert sein. Ein hardwarebasierter Server ist ein Computer, der in ein Rechnernetz eingebunden ist und der so Ressourcen über ein Netzwerk zur Verfügung stellt. Ein softwarebasierter Server hingegen ist ein Programm, das einen spezifischen Service bietet, welcher von anderen Programmen (Clients) lokal oder über ein Netzwerk in Anspruch genommen wird.

Tagset Ein Tagset definiert die Taxonomie, anhand derer Annotationen in einem Projekt erstellt werden. Ein Tagset beinhaltet immer mehrere Tags und ggf. auch Subtags. Ähnlich der Type/Token -Differenz in der Linguistik sind Tags deskriptive Kategorien, wohingegen Annotationen die einzelnen Vorkommnisse dieser Kategorien im Text sind.

TEI Die *Text Encoding Initiative* (TEI) ist ein Konsortium, das gemeinsam einen Standard für die Darstellung von Texten in digitaler Form entwickelt. Die TEI bietet beispielsweise Standards zur Kodierung von gedruckten Werken und zur Auszeichnung von sprachlichen Informationen in maschinenlesbaren Texten (siehe auch XML und Markup).

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.

Webanwendung Eine webbasierte Anwendung ist ein Anwendungsprogramm, welches eine Webseite als Schnittstelle oder Front-End verwendet. Im Gegensatz zu klassischen Desktopanwendungen werden diese nicht lokal auf dem Rechner der Nutzer*innen installiert, sondern können von jedem Computer über einen Webbrowser „online“ genutzt werden. Webanwendungen erfordern daher kein spezielles Betriebssystem.

Wordcloud Eine *Wordcloud* , oder auch Schlagwortwolke, ist eine Form der Informationsvisualisierung, beispielsweise von Worthäufigkeiten in einem Text oder einer Textsammlung. Dabei werden unterschiedlich gewichtete Wörter, wie die häufigsten Wörter, i.d.R. größer oder auf andere Weise hervorgehoben dargestellt. Die horizontale/vertikale Ausrichtung und die Farbe der dargestellten Wörter hat meistens allerdings keinen semantischen Mehrwert.

XML XML steht für *Extensible Markup Language* und ist eine Form von Markup Language, die sowohl computer- als auch menschenlesbar und hochgradig anpassbar ist. Dabei werden Textdateien hierarchisch strukturiert dargestellt und Zusatzinformationen i. d. R. in einer anderen Farbe als der eigentliche (schwarz gedruckte) Text dargestellt. Eine standardisierte Form von XML ist das TEI-XML.

ZIP ZIP steht für ein Dateiformat (zip = engl. Reißverschluss), in welchem mehrere Einzeldateien verlustfrei, komprimiert zusammengefasst werden. ZIP-Dateien werden beim Öffnen entweder automatisch entpackt oder lassen sich per Rechtsklick extrahieren.